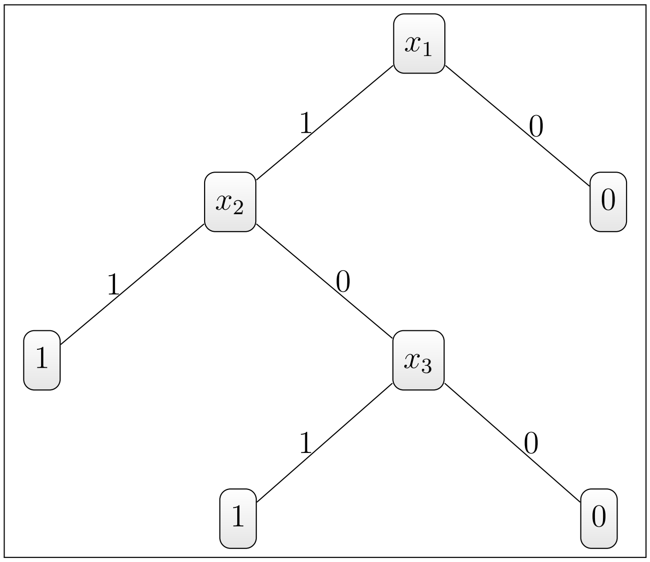
CS 5350/6350: Machine Learining Fall 2017

Homework 1 Solution
UNID: u1143683

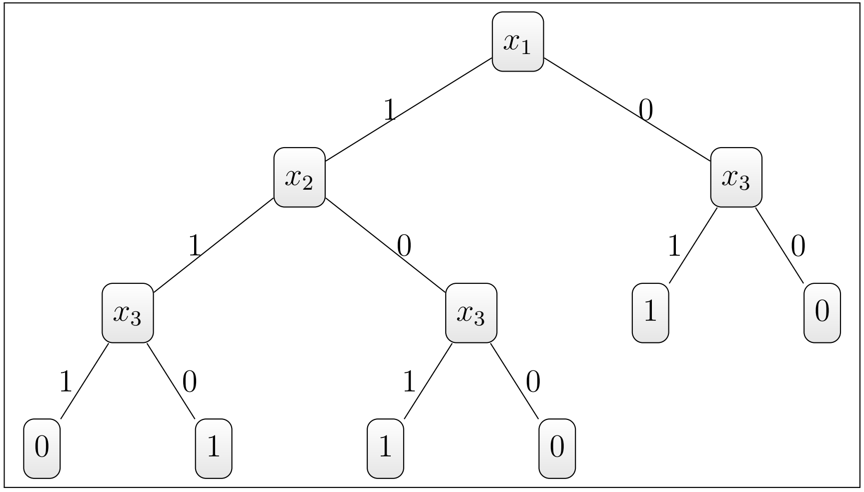
September 13, 2017

1 Decision Tree

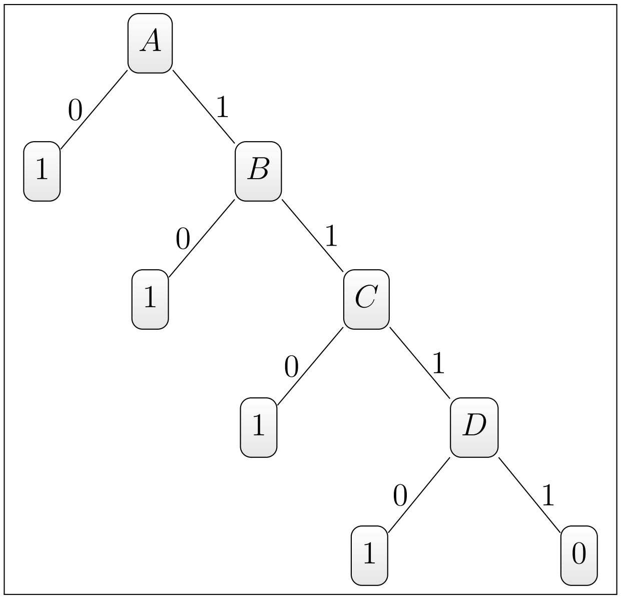
1. (a) $(x_1 \wedge x_2) \vee (x_1 \wedge x_3)$



(b) $(x_1 \vee x_2)$ *xor* x_3



(c) $\neg A \vee \neg B \vee \neg C \vee \neg D$



2. (a)

(b) $S = 9$
$p_+ = 5/9 = 0.56$
$p_- = 4/9 = 0.44$

$Entropy(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$
$Entropy(S) = -0.56 \log_2 0.56 - 0.44 \log_2 0.44$
$Entropy(S) = 0.99$

(c) Information gain of an attribute A is given by,

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

i. Feature $A =$ Technology

For $v = Yes$,
$S_v = 3$
$p_+ = 1/3 \quad p_- = 2/3$
$Entropy(S_v) = 0.918$

For $v = No$,
$S - v = 6$
$Entropy(S_v) = 0.918$

$Gain(S, Technology) = 0.99 - (6/9)0.918 - (3/9)0.918 = 0.071$

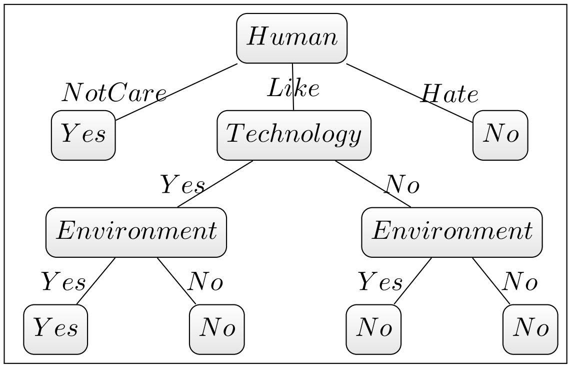
ii. Feature $A =$ Environment
$Gain(S, Environment) = 0.228$

iii. Feature $A =$ Human
$Gain(S, Human) = 0.629$

iv. Feature $A = Distance$
$Gain(S, Distance) = 0.155$

(d) The attribute $Human$ provides the best prediction of target attribute $Invade$ according to the information gain measure. So I will use attribute $Human$ as root node.

(e) Decision Tree using Human Attribute as root node.



(f) Predictions:

Technology	Environment	Human	Distance	Invade(Prediction)	Invade (Given)
Yes	Yes	Like	2	Yes	No
No	No	Hate	3	No	No
Yes	Yes	Lkie	4	Yes	Yes

Accuracy is $2/3 = 66.67\%$

3. (a) Information gain of an attribute A is given by,

$$Gain(S, A) = MajorityError(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} MajorityError(S_v)$$

Total Majority Error, $S = 1 - 5/9 = 0.44$

i. Feature $A =$ Technology

For $v = Yes$,
$S_v = 3$

$p_+ = 1/3 \quad p_- = 2/3$
$MajorityError(S_v) = 0.33$

For $v = No$,
$S - v = 6$
$MajorityError(S_v) = 0.33$

$Gain(S, Technology) = 0.44 - (6/9)0.33 - (3/9)0.33 = 0.106$

ii. Feature $A =$ Environment
$Gain(S, Environment) = 0.217$

iii. Feature $A =$ Human
$Gain(S, Human) = 0.328$

iv. Feature $A = Distance$
$Gain(S, Distance) = 0.106$

The attribute Human has highest information gain with majority error as the measure of impurity.

(b) Highest information gain was obtained for the attribute Human, so it will be selected as the root node for the decision tree. We will obtain a similar tree.

2 Linear Classifier

1. The given table can be represented by the following linear classifier:

$$x_1 + x_4 \geq 1$$

   weight vector $\boldsymbol{w} = [1, 1]$ and *bias* $b = 1$

2. Classification:

x1	x2	x3	x4	o(prediction)	o(given)
0	0	0	1	1	1
0	0	1	1	1	1
0	0	0	0	-1	-1
1	0	1	0	1	1
1	1	0	0	1	1
1	1	1	1	1	1
1	1	1	0	1	1

   Classifier accuracy = 7/7

3. The classifer,

$$x_1 + x_4 \geq 1$$

   weight vector $\boldsymbol{w} = [1, 1]$ and *bias* $b = 1$, fits the given truth table perfectly.

3 Experiments

1.

   (a) The experiment conducted takes in training data and text data as text files and constructs a decision tree using the training data and use the decision tree to classify test data. First the data is processed to seperate the badges and names in two columns. Next features were extracted from the list of names. Multiple choices for features are available here. I have seleceted following features:

      i. First Name is longer than last name
      ii. Middle Name present or not
      iii. Start and end letters of first name are same
      iv. Initial of first name comes alphabetically before initial of last name
      v. Is second letter of first name vowel
      vi. Is first letter of first name vowel
      vii. Is number of letters in first name even
      viii. Is number of letters in last name even

(b) We can consider other features like:
- Is the first letter of their first name a vowel
- Is the number of letters in their first name even
- Are the first letters of both first name and last name same
- Is the sum of the alphabetical positions of letters in first name even

(c)

(d)

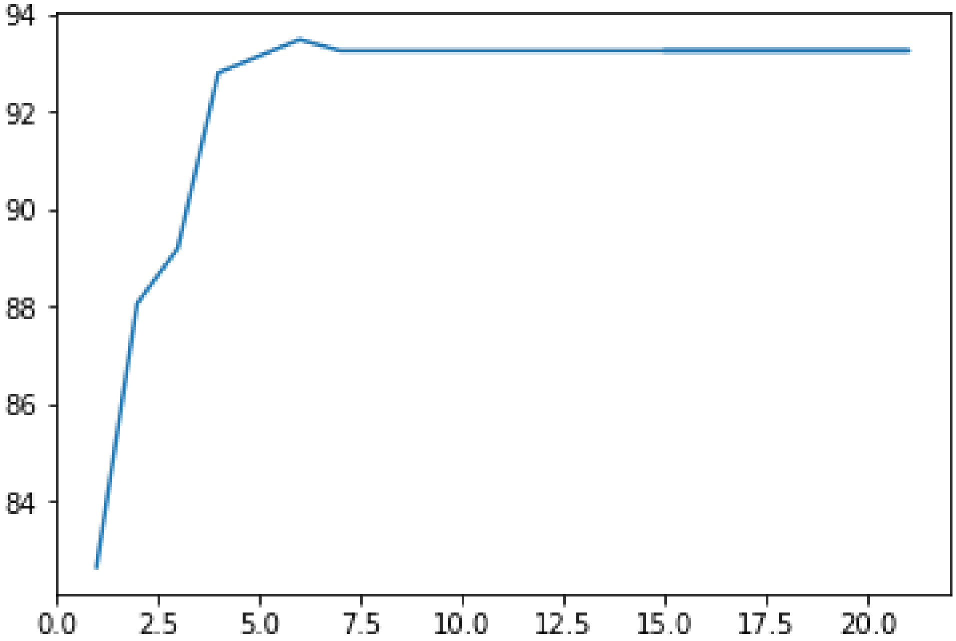
(e)

2.

(a)

(b)

(c) In the above experiment it can be observed that accuracy increases as the function of depth of tree, but the growth tapers off after one point.
This can be seen in the graph below:



Here the X-axis is hyper-parameter depth and Y-axis is accuracy. So yes I think that limiting the depth of a tree is a good idea, because after certain depth the hyper-parameter does not have a substantial effect on improving the accuracy. Trees with lower depths tend to overfit less and have much less complexity.

Experiment Submission Guidelines

1. The report should detail your experiments. For each step, explain in no more than a paragraph or so how your implementation works. You may provide the results for the final step as a table or a graph.

2. *Your code should run on the CADE machines.* You should include a shell script, `run.sh`, that will execute your code in the CADE environment. Your code should produce similar output to what you include in your report.

   You are responsible for ensuring that the grader can execute the code using only the included script. If you are using an esoteric programming language, you should make sure that its runtime is available on CADE.

3. Please do not hand in binary files! We will *not* grade binary submissions.

4 Decision Lists (For CS 6350 students)

Theorem. *Any 1-decision list is a linearly separable function.*

Proof. We will prove the above statement by induction on terms in decision list. Let us assume that any 1-decision list is a linearly seperable function.
Let's consider the base case with only one term x_1. For now we assume x_1 is non negated. We have four different decision lists possible for this scenario. We will select a weight vector $\boldsymbol{w}$ and bias b for each of the case, hence representing the decision list as a linearly separable function.

1. $(x_1, 1), (T, 1)$
   $\boldsymbol{w} = [2]$
   $b = 1$
   We get,
   $$2x_1 + 1 > 0$$

2. $(x_1, 1), (T, 0)$
   $\boldsymbol{w} = [2]$
   $b = -1$
   We get,
   $$2x_1 - 1 > 0$$

3. $(x_1, 0), (T, 1)$
   $\boldsymbol{w} = [-2]$
   $b = 1$
   We get,
   $$-2x_1 + 1 > 0$$

4. $(x_1, 0), (T, 0)$
   $\boldsymbol{w} = [-2]$
   $b = -1$
   We get,
   $$-2x_1 - 1 > 0$$

In case x_1 is negated, we can replace x_1 in the equation with $(1 - x_1)$. So our assumption is true for our base case with one term in decision tree. Now lets assume that any 1-decision list with n terms is also a linearly separable function. Suppose we have a 1-decision list with $n + 1$ terms,

$$f = (a_1, c_1), (a_2, c_2), (a_3, c_3), ..., (a_{n+1}, c_{n+1})$$

By induction hypothesis the decision list

$$(a_2, c_2), (a_3, c_3), ..., (a_{n+1}, c_{n+1})$$

of length n is a linearly separable function. Let it be represeted by weight vector $\boldsymbol{w}$ and bias b. Let $||w||_1 = \sum_{i=1}^{n} |w_i|$ be the 1-norm of $\boldsymbol{w}$. There are now four different values the first

term (a_1, c_1) can take as follows: $(x_1, 1), (x_1, 0), (\bar{x_1}, 1), (\bar{x_1}, 0)$.
Let $e_1 = [1, 0, 0, 0....0]$ be a vector and

$$A = ||w|| + |b| + 1 \tag{1}$$

Now we claim that decision list f is a linearly separable function represented by weight vector w' and bias b' where,

$$w' = w + Ae_1,\ b' = b$$

$$w' = w - Ae_1,\ b' = b$$

$$w' = w - Ae_1,\ b' = b - A$$

$$w' = w + Ae_1,\ b' = b - A$$

To verify our claim let us consider case where (a_1, c_1) takes the form $(x_1, 1)$. For $x \in \{0, 1\}^n$.

$$\langle w', x \rangle = \langle w + Ae_1, x \rangle = \langle w, x \rangle + Ax_1$$

Now,

$$\langle w', x \rangle \geq b' = b$$

$$\langle w, x \rangle + Ax_1 \geq b$$

If $x_1 = 1$ we have, $\langle w, x \rangle - A \geq b$

$$\langle w, x \rangle + ||w|| + |b| + 1 \geq b$$

$$\langle w, x \rangle \geq -||w|| - 1$$

Now, $||w|| + 1 > ||w||$

$$\langle w, x \rangle > -||w||$$

Above statement is true for $x \in \{1, 0\}^n$, so output of the function is 1. Now consider $x_1 = 0$, we have

$$\langle w, x \rangle \geq b$$

By inductive hypothesis the decision list,

$$f' = (a_2, c_2), (a_3, c_3), ..., (a_{n+1}, c_{n+1})$$

is linearly separable function with weight vector $\boldsymbol{w}$ and bias b. Hence output of the linear separable equation is 1 only when output of decision list f' is 1, which is how f calculates its output when $x_1 = 0$. we can say f is indeed a linearly separable function. Other claims can be verified in a similar manners.

□

The idea for substitution in (1) was as seen in 'CDAM Research Report LSE-CDAM-2002-11' titled 'Threshold Decision Lists' by Martin Anthony. We make this substituion as a claim and go on to prove it.